Assessing the Effect of Natural Selection in Pinpointing Ancestry along the Genome of an Individual of Mixed Ancestry

Justin USHIZE RUTIKANGA (justinushize@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Dr. Emile Chimusa Rugamika
University of Cape Town, South Africa

22 May 2014

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

Human migrations have played an important role in shaping the genetic diversity of human populations. Therefore, the analysis of the genetic data of populations having been formed from mixing of two or more parental populations (which are genetically distinct) provides important insights into medical genetics and population history. Such analyses have been used to identify novel disease genes, to understand recombination rate variation and to detect recent selection events. The utility of such studies crucially depends on accurate and unbiased estimation of the ancestry at every genomic locus in admixed populations, since the chromosomes of individuals that have ancestry from multiple source populations are mosaics of segments originating from each population. Although various methods have been proposed, a major limitation is that pinpointing ancestry along the genome of a complex multi-way admixed population such as the South African Coloured population is currently an unsolved problem. Existing methods may attain high accuracy on average but may suffer from spurious deviations in average local ancestry at particular regions (e.g.regions in which the modelled ancestral population is unusually different from the true ancestral population due to the historical action of natural selection). These spurious deviations would be present in both affected and unaffected individuals, and would lead to spurious mapping of genes underlying ethnic difference in disease risk. To address these challenges, this project aims to discuss a variety of hidden Markov models of switches in ancestry between consecutive windows of loci and fit model parameters using the model-Markov chain Monte Carlo sampling, the Expectation Maximization algorithm, or a Classification Expectation Maximization algorithm. We additionally, aim to assess the effect of natural selection in different implemented hidden Markov Models using the Tuberculosis case-control data of the South African Coloured population.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Justin Ushize Rutikanga, 22 May 2014

Contents

1. Introduction and Background

1.1 Population Genetics of Admixture

1.1.1 Genetics of Admixture. The analysis of genetic data of populations have been formed from mixing of two or more parental populations to provide important insights into medical genetics and population history (Baran et al., 2012; Redden et al., 2006; Qin et al., 2010). Furthermore, high-resolution ancestry inference from genome-wide genotype data of the mixed individuals, forms an essential analysis step in medical genetics in the identification of disease gene association and is a critical process in personalized drug therapy applications and pharmacogenomics (Pasaniuc et al., 2013; Baran et al., 2012; Seldin et al., 2011). Among several factors, human migrations and the mixture of individuals isolated during a long period of time, have been shown to be the source of the genetic diversity of human populations over the World. A critical question is why the populations around the world are genetically different, and why they have difference in disease risk or drugs response?

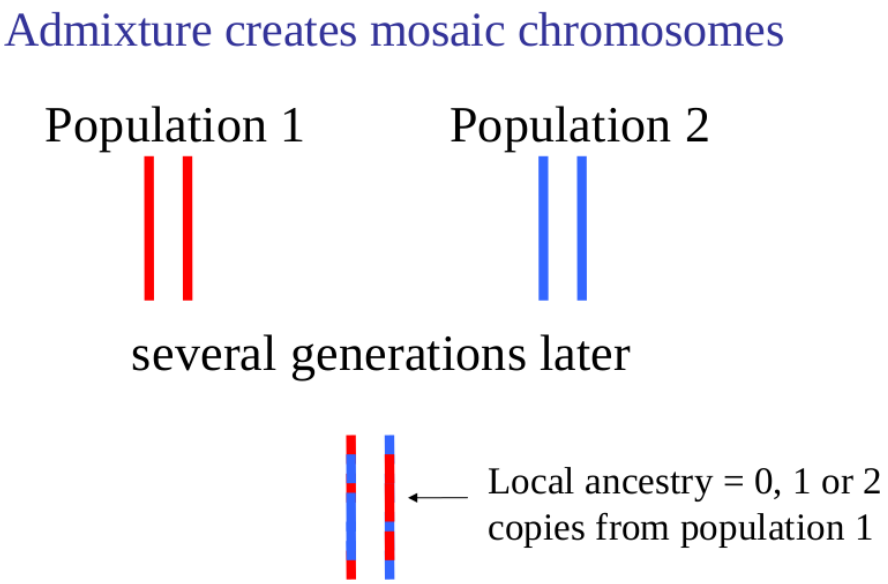


Figure 1.1: Two populations admixture event(Alkes, 2010).

Throughout human history, contacts between two or more previously isolated populations are mostly due to different population migrations, colonization waves, or forced displacements due to many reasons such as ecology, climate, agriculture and hunting. Most of these human contacts or admixture processes (Figure 1.1) have been influenced by sociocultural laws on inter-marriage. It has been shown that the mixture of previously isolated populations, results in admixed populations (Figure 1.1) that benefit from several genetic advantages such as increased genetic variation, the creation of novel genotypes and the masking of deleterious mutations (McKeigue, 2005; Halder and Shriver, 2003). These admixture benefits are thought to play an important role in biological invasions (Verhoeven et al., 2010), in assessing patterns of migration and genetic structure (Pritchard et al., 2002), in detecting natural selection (Tang et al., 2006; Lohmueller et al., 2011), and in providing valuable baseline data for subsequent analysis of disease association, particularly identifying phenotypically relevant genes through admixture-mapping strategies (McKeigue, 2005; Halder and Shriver, 2003; Reich et al., 2005; Stephens et al., 2005; Seldin et al., 2011). A specific location in the chromosomal region of the admixed individual may inherit 0,

1, or 2 copies of a particular ancestry allele (Figure 1.1). Inferring an individual's local ancestry, or their number of copies of each ancestry at each location in the genome, also has important applications in disease mapping and in understanding human history. As the genomes of individuals from admixed populations consist of chromosomal segments of different ancestry, a specific location in the genome may contain 0, 1, or 2 copies (local specific ancestry or local ancestry) from a particular ancestral population. It has been shown that the inference of an individual's local ancestry have a wide range of applications from disease mapping to learning about history (Price et al., 2009; Sankararaman et al., 2008).

1.2 Challenges and Opportunities in Pinpointing the Mixed Individual's Ancestry

An accurate and unbiased estimation of the ancestry at every genomic region in admixed populations may potentially provide crucial insights into identifying disease gene associations, and provide information on the timing of the ancient or recent admixture event itself in any admixed population (Seldin et al., 2011). Because of the importance of the inference of ancestry in both understanding population history and disease scoring statistics, number of methods for inferring local ancestry have been proposed (Pasaniuc et al., 2009; Baran et al., 2012; Lawson et al., 2010; Rodriguez et al., 2012; Henn et al., 2012; Churchhouse and Marchini, 2012) and have been shown to be very accurate in admixed populations from just two source of ancestry such as African-Americans (Seldin et al., 2011). Most previous methods of ancestry inference are based on hidden Markov models (HMM), where the hidden states correspond to the unknown ancestral populations that generate the observed genotypes.

Methods have improved for both ancient and recent admixture events between more closely related populations, but challenges remain in accurately inferring local ancestry for multi-way admixed populations and accounting for admixed parental populations (Relethford and Harding, 2001). Nevertheless, high-throughput genotyping or sequencing and new methods to infer local ancestry can allow for joint admixture and association analysis. This may benefit association mapping in admixed populations by eliminating the effects of confounding due to variation in ancestry (Price et al., 2009; Baran et al., 2012).

However, in the last few years a statistical model has been produced in order to find the probable ancestral origin of chromosomal segments and to understand the mosaic structure of the genome of admixed populations. A specific location in the genome may inherit 0, 1 or 2 copies of a particular ancestry. Therefore, inferring an individual's local ancestry at each location in the genome has critical application in detecting genes that underlie ethnic differences in a specific heritable disease risk.

Different method for inferring local ancestry have been developed and these approaches can be clustered to the following categories (Alkes et al., 2009).

(1) Haplotype-based inference of local ancestry inference methods such as HAPMIX (Price et al., 2009) SPECTRUM (Kyung-Ah and Eric, 2007) and HAPAA (Sundquist et al., 2008) make use of all SNPs of the genomes of the admixed populations. The Haplotype-based inference makes use of Hidden Markov models (HMMs) based on the population-specific allele frequency profiles. This approach is known to be accurate when using two-way admixed populations (Alkes et al., 2009; Kyung-Ah and Eric, 2007).

(2) Principal Component Analysis-based. Local ancestry assigned a new method such as PCADMIX

(Brenna et al., 2012) relies on Principal Components Analysi (PCA) to quantify the information that each SNP contributes to differentiating the ancestry of a genomic region of an admixed population. PCADMIX is one of the multiple locus-specific ancestry assignment methods which requires thinning genotype data sets to remove alleles in high linkage disequilibrium between populations (Brenna et al., 2012).

(3) Imputation-based approach including ALLOY is a novel locus-specific ancestry inference method that enables the incorporation of complex models for linkage disequilibrium in the ancestral populations. This method applies a factorial hidden Markov model to capture the parallel process producing the maternal and paternal admixed haplotypes. In addition, these method models background LD in ancestral populations via an inhomogeneous variable length Markov chain (Jesse et al., 2013).

(4) Window-based approach, is the most ancient method that explores window-based techniques, in which a simple ancestral composition is assumed to occur within a window example include LAMP and WINPOP. In particular WINPOP works in windows side along the chromosomal region, however, it assumes at most one recent recombination in each window. In order to find the infers local ancestry estimate by partition of genome into overlapping and contiguous of SNPs. However, it optimizes the probability in the new model and combine the solution by throwing a majority of SNPs (Gad et al., 2009).

In the former years, the interest in admixed populations is increasing, in 1988, Chakraborty ad Weiss described a novel approach for mapping disease gene association, using the chromosomal's ancestry of the admixed population. Linkage disequilibrium is created when two ethnically different populations are mixed; and it is useful for mapping disease genes, for diseases that show high level of prevalences among these two parental populations. For example the hypertension, lung and prostate cancer among Africa-Americans and the diabetes and obesity among Hispanics and African-Americans(Bogdan et al., 2009). Those diseases show high difference in prevalence between their parental populations (ancestral population include African and European populations). In addition, admixed populations are an important resource that can and should be employed to examine the complex diseases (Relethford and Harding, 2001). A basic to this application is better understanding of the admixture proportion and kinetics of the admixture process (Esteban et al., 1998).

The inference of local ancestry using admixed populations of more than three ancestries is currently an unsolved problem in that existing methods may attain high accuracy on average but may suffer from spurious deviations in average local ancestry at particular regions (e.g. regions in which the modelled ancestral population is unusually different from the true ancestral population due to the historical action of natural selection). These spurious deviations would lead to wrong associations when mapping disease genes underlying ethnic difference(Chimusa et al., 2012).

1.3 Motivation and Overview of the Essay

Since human migrations (Figure 1.2) have played an important role in shaping the genetic diversity of human populations, analysing the genetic data of populations have been formed from mixing of two or more parental populations may provides important insights into medical genetics, individualized medicine and population history. Satisfactory results and interpretations rely on accurate and unbiased estimation of the ancestry at every genetic locus in admixed populations. In fact, the chromosomes of individuals that have ancestry from multiple source populations are as mosaics of segments originating from each population. Although various methods have been proposed, however a major limitation is

that pinpointing ancestry along the genome of a complex multi-way admixed populations such as the South African Coloured and Latino populations (Pasaniuc et al., 2013) is currently an unsolved problem (Pasaniuc et al., 2013; Chimusa et al., 2012). Existing methods may attain high accuracy on average local ancestry, but at a particular regions may obtain a wrong deviations in average local ancestry for case versus control (e.g. regions in which the modelled ancestral population is unusually different from the true ancestral population due to the historical action of natural selection). These spurious deviations would be present in both affected and unaffected individuals, and would lead to spurious mapping genes underlying ethnic difference in disease risk. To address these challenges, this project aims to discuss various of hidden Markov models of switches in ancestry between consecutive windows of loci and fit model parameters using the model-Markov chain Monte Carlo sampling, the Expectation Maximization algorithm, or a Classification Expectation Maximization algorithm.

The central premise of this essay is concerned with the discussion of variety of hidden Markov models of switches in ancestry between consecutive windows of loci and the fit model parameters using the model-Markov chain Monte Carlo sampling; forward-backward algorithm and the Expectation Maximization algorithm. Furthermore, we aim to assess the effect of natural selection in different implemented hidden Markov Models using the Tuberculosis case-control data of the South African Coloured population. Through the real data of the South Africa Coloured population, we aim to assess spurious deviations in the average local ancestry using different implemented local ancestry inference methods.

This project involves two main axes of investigation:

(1) Mathematical discussion of various hidden Markov models in pinpointing ancestry along the genome of an admixed individual.

(2) Analysis of the Tuberculosis case-control data of the South Africa Coloured (SAC) population, to estimate local ancestry using different approaches of local ancestry inferences and assess their level of producing spurious deviation in average local ancestry between cases and controls (Differing region of natural selection in cases and controls).

1.4 Human History and Variation

1.4.1 Overview. The Human's origin implies that we, as individuals, are all related to one another by differing degrees, thus it is critical to be clear about the frame of reference being used in discussions of "ancestry" and relationship. For example, because of recombination and admixture events, each segment of the genome has its own ancestral history, and various segments of an individual's genome may have ancestral histories (Relethford and Harding, 2001). By studying human migration with DNA, Y-chromosome phylogeography, palaeoanthropological and palaeoclimatological data, researchers argue that the evolution of modern human diversity arose approximately 100 thousand years ago. This started with a tiny population of about $1,000$ individuals from Africa, who migrated to other parts of the world (L.Luca and marcus, 2003).

The most important concerns of the human diversity are based on understanding the consequence of the past human migrations (Figure 1.2), the causes of human diversity in the world today and the related evolutionary history that generated the latter, through mathematical modelling of complex patterns of geographic genetic diversity (Figure 1.2). These patterns of geographic genetic diversity include mutation, natural selection, genetic drift and gene flow that change within and between populations. A number of studies have examined how genetic variation is distributed geographically, and have established that human population differences are mainly due to the presence of low-frequency alleles that have

not diffused far from their geographic place of origin. In addition, a recent study by Rosenberg and colleagues demonstrated that the worldwide human genetic variation within human populations is larger $(93-95\%)$ than that seen between populations $(5-7\%)$ (Rosenberg and Pritchard, 2008; Rosenberg et al., 2003), suggesting that classification of the human species according to racial or continental lines appears to be inappropriate descriptors of the distribution of human genetic variation (Tishkoff and Kidd, 2004).

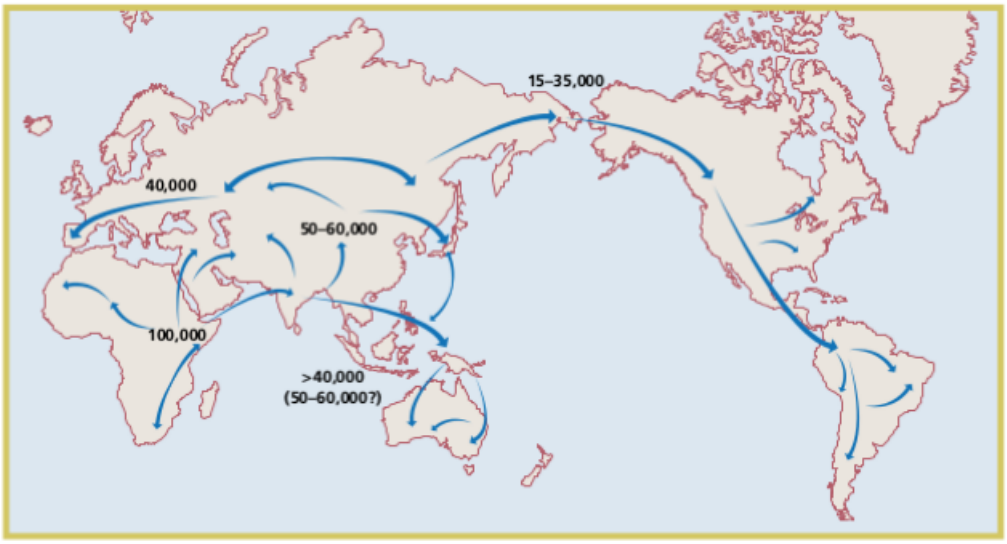


Figure 1.2: The migration of modern Homo sapiens. The scheme outlined above begins with a radiation from East Africa to the rest of Africa about 100 kya, followed by an expansion from the same area to Asia, probably by two routes (southern and northern) between 60 and 40 kya. Oceania, Europe and America were settled in from Asia, in that order (L.Luca and marcus, 2003).

1.4.2 Factors of Genetic Diversity. The investigation of human genetic diversity requires an understanding of different factors of genetic evolution such as:

(1) Natural selection: is known as the difference between alleles in determining the probability for the individual survival, and continuing to maintain the reproduction of more offspring. However, it removes negative alleles transmitted to the next generation to sustainer it in that environment.(Relethford and Harding, 2001; L.Luca and marcus, 2003)

(2) Mutations which are the change of DAN occurred by error during the replication of DAN, which introduced new alleles into the population. These errors can be caused by the physical and chemical phenomena.

(3) Genetic drift: is the difference among the allele frequencies from a generation to the next one as a result of random sampling and has highly frequent in small population(Relethford and Harding, 2001).

(4) Recombination: as the exchange of genetic information between nucleotide sequences. It is an important process that influences biological evolution at many different levels. However, recombination breaks down linkage disequilibrium and, as result, the characterization of the recombination is essential for gene mapping, quantitative trait loci, and association studies (Relethford and Harding, 2001). In addition, the recombination has a significant impact on the evolution of several human pathogens and consequently on their clinical treatment and prevention.

(5) Population bottlenecks occur when a population's size is reduced for at least one generation. Which caused by the promote inbreeding depression and hinder a population's ability to adapt to environmental changes. Such causes affect the reduction of the genetic variation in populations and they are expected to increase a population's risk of extinction (Mary et al., 2010; F.J and C.H.kuo, 2003).

1.5 Quantifying Genetic Variation

1.5.1 Genetic Distance, (F_{st}). From the previous studies, the population differentiation was caused by the genetics variation among the population. By comparing the level of genetic variation is arranged in range from 0 up to 100% in population. However, the population characteristics such as the genetic distance between haplotypes and diversity of the haplotypes are shared among the population (Christopher et al., 2011). Greater genetic distance between populations, less the breeding is between them and the are more isolated from one another and vice versa.

The literature was surveyed to quantify the human genetic variation within and between human population using Wright's F_{st} statistic (Weir and Cockerham, 1984; Weir, 2008). The F_{st} of overall genetic variants is given as,

$$F_{st} = \frac{\sum_{i=1}^{L} p_i^*(1-p_i^*) - F_i}{\sum_{i=1}^{L} p_i^*(1-p_i^*)}, \tag{1.5.1}$$

where p_i^* is the average allele frequency of the i^{th} allele, L is the number of locus, and F_i is the value of F_{st} for each allele. Thus,

$$F_i = \frac{\sum_{j=1}^{2} (p_i^j - p_i^*)^2}{p_i^*(1-p_i^*)}, \tag{1.5.2}$$

where p_i^j is the frequency of the i^{th} allele in the population j.

1.5.2 Nature and Measures of Linkage Disequilibrium. When genotypes at two loci are dependent one to an another, this phenomenon is known as Linkage disequilibrium (LD). In fact,linkage disequilibrium (LD) is the non random association of alleles at two or more loci that may or may not be on the same chromosome(Sagiv et al., 2003). Let's consider populations with different prevalence for certain disease and with different alleles frequencies. If these populations have been isolated for a long time and then they are admixed the resulting hybrid chromosomes are transmitted to the offspring and this process continues into the next generations(McKeigue, 2005; Rosenberg and Pritchard, 2008). Hence, even unlinked genetic markers can generated the linkage disequilibrium because of that admixture. When the genetics markers influence the traits, this has an impact in the divergence of the SNPs frequencies.

LD varies among different populations and genome regions; and between pairs of the genetic markers in close proximity (Reich et al., 2005). In addition, there exist different factors, that contribute in the differentiation of LD, such as genetics drift, admixture and inbreeding which are known as population specific. Moreover, there are other factors that play important role in the contribution to extend and distribution of linkage disequilibrium such as recombination rate, gene conversion and natural selection. They are specific to the genomic region(Reich et al., 2005; Weir, 2008; Kristin et al., 2002). Thus, the strength relationship between linkage disequilibrium and the physical position between genetic loci tend to be distort, because of the involvement of the demographic aspects in the populations. Therefore, this

shows that it is possible to restrict the genetic interval around the disease locus by identifying linkage disequilibrium between near by genetic markers and the disease locus, if the most affected individuals in population share the same mutant allele at a causal locus.

The most popular measures of linkage disequilibrium is the r^2 coefficient of correlation in equation 1.5.5 which determines the level of correlation of two given different loci with two alleles.Let's consider two different loci A and B with two alleles A, a and B, b at each genetic locus, respectively. The measure of linkage disequilibrium is given by

$$D = f_{AB} - f_A f_B, \tag{1.5.3}$$

where f_{AB} is the observed frequency of the haplotype of alleles A and B, $f_A f_B$ is the expected haplotype frequency in the absence of linkage disequilibrium and f_A, f_B are the alleles frequency of the allele A and B, respectively. There exist alternative measures based on the measure of D, since these measures have different properties and quantify different things, it might be difficult to compare different reports on the extent of linkage disequilibrium(DF.Conrad et al., 2010). In addition, there are three different types of linkage disequilibrium in human population:

(1) Mixture linkage disequilibrium which is due to the population admixture between unlinked genetic markers and as known to be the main source of inflated type I error in case-control in association studies.

(2) Admixture linkage disequilibrium which occurs when considerable chromosomal segments are transmitted from a particular ancestral population. It provides the necessary basis for conducting association studies.

(3) Background linkage disequilibrium which exists within ancestral population because of correlation among polymorphisms is very low and is viewed as the main subject of case control association studies (Montana and Pritchard, 2004).

Since the measure D given in equation 1.5.3 depend on the alleles frequency and where it is not commonly used to measure the strength of linkage disequilibrium. The normalized measure, D' of D and r^2 are known as the most popular measure of linkage disequilibrium (Evans and Cardon, 2005). The normalized D' is obtained by dividing D by it maximum possible value of the given alleles frequencies at the two genetic loci with alleles A and B, respectively.

$$\begin{cases} D' = |\frac{D}{max(f_A(1-f_B), f_B(1-f_A))}| & \text{where } D < 0, \\ D' = |\frac{D}{min(f_A f_B, (1-f_A)(1-f_B))}| & \text{where } D > 0. \end{cases} \tag{1.5.4}$$

Where $D' = 1$ if and only if two SNPs have not been separated by recombination during the history of the sample and there is a complete linkage disequilibrium. The value of $D' < 1$ indicates that the complete ancestral linkage disequilibrium has been perturbed and for $D' > 1$ there is no clear interpretation(Evans and Cardon, 2005).

The measure r^2 is a complementary of the value D' and has recently emerged as the measure of the quantifying and comparing linkage disequilibrium in the context of the mapping correlation. It is like the Pearson correlation coefficient of the alleles at two loci and it is obtained by dividing D^2 by the product of the four alleles frequencies at two genetic loci.

$$r^2 = \frac{D^2}{f_A f_B f_a f_b}. \tag{1.5.5}$$

For $r^2 = 1$ is known as the perfect linkage disequilibrium and it occurs only if the markers have not been separated by recombination and have the same alleles frequency (Ranajit and Kennethe, 1988).

1.5.3 Population Structure. Population structure refers to the genetic differentiation between population due to the geographic ancestry. Population structure is a model-based clustering approach includes STRUCTURE (Pritchard et al., 2002), FRAPPE (Liu et al., 2013) and ADMIXTURE (Liu et al., 2013) and Principal components analysis (PCA) methods to classify the genome-wide ancestry contribution in a given population (homogeneous or admixed). In model-based clustering, the fractional ancestry of the individual from different populations must be known and the result is sensitive to the number of population clusters but (PCA) is not sensitive to number of principal components.

Model-based clustering is used when we are interested in finding which population belong to an individual(s) and what fraction of ancestry an individual inherits from that population. These two tasks can be determined in two cases, that is when allele frequencies are known and when allele frequencies are not known.

Consider $Y(\ y \in Y)$, populations with $X(\ x \in X)$, Single Nucleotide Polymorphism(SNPs) and allele frequency f_{yx} for each SNP X in each population Y. We also denote q_x to be the observed genotype that is counted for each of the X SNPs in the individual z. Now let us consider the first case when the frequencies of the alleles are known. If we are interested in finding which population y an individual z belongs to, when we have all the above information, then it will be computed with the expression,

$$P(DATA \mid z \sim \text{population y}) \simeq \prod_{x=1}^{X} (f_{yx})^{q_x} (1 - f_{yx})^{2-q_x} \tag{1.5.6}$$

Again to find the fraction of ancestry (β_y) that an individual z inherits from each populations when the frequencies are known is given by the expression,

$$P(DATA \mid z \sim \beta_1, ..., \beta_Y) \ \text{ is proportional to } \ \prod_{x=1}^{X} \left(\sum_{y=1}^{Y} \beta_y f_{xy} \right)^{q_x} \left(\sum_{y=1}^{Y} \beta_y (1 - f_{yx}) \right)^{2-q_x} \tag{1.5.7}$$

The values of β_y that maximizes the above expression can be computed using Expectation maximization algorithm or other sampling methods such as Monte Carlo. Considering the other case when the frequencies of the alleles are not known, to find the fraction of ancestry (β_y) that an individual inherits from each populations is given by

$$P(DATA | z \sim \beta_1, ..., \beta_Y; f_{yx}) \text{ is proportional to } \prod_{x=1}^{X} \left(\sum_{y=1}^{Y} \beta_y f_{yx} \right)^{q_x} \left(\sum_{y=1}^{Y} \beta_y (1 - f_{yx}) \right)^{2-q_x} \tag{1.5.8}$$

The value of β_y and f_{xy} that maximize the expression can be computed using any sampling methods.

Let consider the case where the allele frequencies are unknown and in this case we are interested in finding the fractional ancestries for many individuals z_j across Y populations. We denote the observed genotype counts for X SNPs in many individuals z_i by q_{jx} and β_{jy} as the fractional ancestries of individual z_j from each of the y populations where the sum of the fractional ancestries that an individual z_j has in all the y populations should be 1, $(ie \sum_y \beta_{jy} = 1)$. Hence the most likely values for β_{jy}, that is finding each fractional ancestry across populations for each individual is given by the expression

$$P(DATA|z \sim \beta_{i1}, ..., \beta_{jY} \text{for each j; } f_{yx}) \text{ is proportional to } \prod_{j=1}^{J} \prod_{x=1}^{X} \left(\sum_{y=1}^{Y} \beta_{jy} f_{yx} \right)^{q_{jx}} \left(\sum_{y=1}^{Y} \beta_{jy}(1 - f_{yx}) \right)^{2-q_{jx}} \tag{1.5.9}$$

where we compute the values of β_{jy} and f_{yx} that maximize the expression.

1.5.4 Local ancestry. The genome of individuals from the admixed populations formed of the chromosomal segment of various ancestries such as the genome of the South Africa Coloured individual who may contained the segment of the African, European, Chines, Indiana ancestries.This implies that it indicates that a local ancestry at a specific region in the genome may inherit 0, 1, 2 copies of African ancestry as an example (Alkes, 2010). Hence, inferring an individual's local ancestry, the number of copies of each ancestry at each region in the genome has a particular applications in testing genes that handle ethnic difference in particular inherited disease risk such as asthma, diabetes, obesity, tuberculosis, heart disease, hypertension and different cancers. In order to identify local ancestry of chromosomal segment of different ancestries has many application in association studies and disease mapping to understand about history of population events (Brenna et al., 2012). The different methods have been developed for inferring local ancestry, and these methods are divided into the following groups:

(a) Overlapping windows based inference on the all genome data such as LAMP, WINPOP. This methods inferring the local ancestry by dividing the genome into overlapping, and contiguous windows of SNPs. Then, it approximates the likelihood for each windows and make together the solution by casting a majority vote form each SNP (Gad et al., 2009)

(b) Haplotype-based inference of local ancestry such as SABER, HAPAA which makes important role of all SNPs of the genome of the admixed populations. The haplotype-based inference applied a Hidden Markov Model (HMM) regarding on the population allele frequency profiles (Kyung-Ah and Eric, 2007).

Therefore, let us assume m be the be genome-wide ancestry and α denoted as number of generation from the starting on the admixture. Let X_i be the hidden state of the individual at markers i along the genome.

Where $X_i \in \{0, 1, 2\}$, with initial probabilities

$$\begin{aligned} P(X_0 = 0) &= (1 - m)^2, \\ P(X_0 = 1) &= 2m(1 - m), \\ P(X_0 = 2) &= m^2. \end{aligned} \tag{1.5.10}$$

The transition probabilities are given by,

$$\begin{aligned}
P(X_i = 0|X_{i-1} = 0) &= e^{-2d\alpha} + 2e^{-\alpha d}(1-m) + (1-e^{-\alpha})^2(1-m)^2, \\
P(X_i = 1|X_{i-1} = 0) &= 2e^{-\alpha d}(1-m) + 2(1-e^{-\alpha d})^2(1-m)m, \\
P(X_i = 1|X_{i-1} = 0) &= (1-e^{-\alpha d})^2 m^2, \\
P(X_i = 0|X_{i-1} = 1) &= 2e^{-\alpha d}(1-e^{-\alpha d})(1–m) + (1-e^{-\alpha d})^2(1-m)^2, \\
P(X_i = 1|X_{i-1} = 1) &= e^{-2\alpha d} + e^{-\alpha d}(1-e^{-\alpha d}) + (1-e^{-\alpha d})^2 2m(1–m), \\
P(X_i = 2|X_{i-1} = 1) &= e^{-\alpha d}(1-e^{-\alpha d}) + (1-e^{-\alpha d})^2 m^2, \\
P(X_i = 0|X_{i-1} = 2) &= (1-e^{-\alpha d})(1-m)^2, \\
P(X_i = 1|X_{i-1} = 2) &= 2e^{-2\alpha d}(1-e^{-\alpha d})(1-m) + (1-e^{-\alpha d})^2 2m(1–m), \\
P(X_i = 2|X_{i-1} = 2) &= e^{-2\alpha d}(1-e^{-\alpha d}) + (1-e^{-\alpha d})^2 m^2.
\end{aligned} \tag{1.5.11}$$

Emission probabilities Let p and q be the genotype frequencies of a marker i in two admixed populations for example African and European. the emission probabilities are given by,

$$\begin{aligned}
P(g_i = 0|X_i = 0) &= (1-p)^2, \\
P(g_i = 1|X_i = 0) &= 2p(1-p), \\
P(g_i = 1|X_i = 0) &= p^2, \\
P(g_i = 0|X_i = 1) &= (1-p)(1-q), \\
P(g_i = 1|X_i = 1) &= p(1-q) + q(1-p), \\
P(g_i = 2|X_i = 1) &= pq, \\
P(g_i = 0|X_i = 2) &= W(1-q)^2, \\
P(g_i = 1|X_i = 2) &= 2q(1-q), \\
P(g_i = 2|X_i = 2) &= q^2.
\end{aligned} \tag{1.5.12}$$

Thus, we can apply forward-backward algorithm to infer $P(X_i|\text{genotypes})$ which is the local ancestry inference.

In next chapter, we will discuss in details different approaches of hidden Markov used to infer the local ancestry of admixed populations.

2. Probabilistic Approach

2.1 Introduction

The chromosomes of admixed individuals who have a recent ancestry from two or more genetically diverged populations may considered as being composed of segment of different ancestry(Churchhouse and Marchini, 2012). That divergence may provide information on the historical population events, and it is helpful for the detection of single nucleotide polymorphisms (SNPs) associated with diseases through association studies and admixture mapping (Baran et al., 2012).

However, the one of the key for understanding the genetic variation in a population is the local ancestry inference(Baran et al., 2012). There exist different methods, all based on the variety of hidden Markov models that researchers used to model the ancestral origin of chromosomal segment in admixed individuals (Rodriguez et al., 2012). However, the improvement of these methods was motivated by various applications, like studying migration association and improving admixture mapping for both disease-gene mapping as well as personalised drug therapy (Rodriguez et al., 2012). Hence, the strength of these disease scoring statistics methods rely on the ability to accurately guess the ancestry of origin along the chromosomes of the admixed individuals.

This chapter is concerned with discussing different methods of hidden Markov models and how the posterior distributions are approximated using different algorithm including the forward-backward and the Expectation Maximization (EM) to estimate the local ancestry.

2.2 Reviews of Mathematical Modes for Local Ancestry Inferences

2.2.1 Classical hidden Markov model. The designation of an ancestry chromosomal segment of the different continental ancestry in admixed populations is an important problem, with a wide range of application from disease mapping to understanding human history. However, in previous years, a statistical model has been developed in order to detect the exact ancestral origin of chromosomal segments and help us to understand the mosaic structure of the genome of admixed populations.

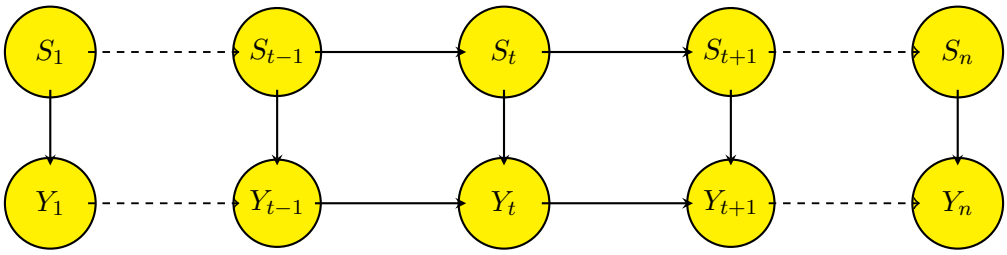


Figure 2.1: The graph shows standard Hidden Markov model.

By considering the Hidden Markov model(HMM) which is a statistical model used to show the probability distribution of the alleles sequence of unobserved state as unknown ancestry for a given observation state (Jesse et al., 2013). Let's consider a sequence $\{S_t\}_{t=1}^n$ of n genetics markers loci along the chromosome and related sequence $\{Y_t\}_{t=1}^n$ of the hidden number of ancestral alleles at the corresponding genetic markers loci. The sequences $\{S_t\}$ and $\{Y_t\}$ are independent and the model works with the following

assumption: S_t is independent of all other observations given Y_t, and that Y_t is independent of $Y_1 \cdots, Y_{t-2}$ given Y_{t-1}.

Thus, we can denote $\{S_t, Y_t\}_{t=1}^{n}$ as the hidden Markov model as shown in (Figure 2.1). From the hidden parameters Markov model, we can derive some parameters to fit the coalescence model (mutation, natural selection, genetic drift) through the joint distribution, the transition probability and the emission probability as follows:

Let us consider the random variables $S_1, \cdots, S_n$ and $Y_1, \cdots, Y_n$. Now we can find the joint distribution from the (Figure 2.1),

$$P(S_1, \cdots, S_n, Y_1, \cdots, Y_n) = P(S_1)P(Y_1|S_1)\prod_{k=2}^{n}(S_k|S_{k-1})P(Y_k|S_k). \tag{2.2.1}$$

The transmission probability T can be governed by

$$T(i,j) = P(S_{k+1} = j|S_k = i) \quad i,j,k \in \{1, \cdots, n\}. \tag{2.2.2}$$

Let ϵ denote the emission probability which can be expressed in term of the probability density function

$$\epsilon_{(i)}(s) = \rho(S_k|Y_k = i), \quad i \in \{1, \cdots, n\}, \tag{2.2.3}$$

$$\epsilon_i(s) = P(Y_k = s|S_k = i). \tag{2.2.4}$$

Let π be denoted the initial distribution such that

$$\pi(i) = P(Y_1 = i), \quad i \in \{1, \cdots, n\}. \tag{2.2.5}$$

Therefore, by substituting equations 2.2.4 and 2.2.3 in equation 2.2.5 we obtain the joint distribution expressed in terms of the initial distribution, the transmission and emission probability as follows,

$$P(S_1, \cdots, S_n, Y_1, \cdots, Y_n) = \pi(i)\epsilon_i(s)\prod_{k=2}^{n}T(Y_{k-1}, Y_k)\epsilon_{y_k}(s_k). \tag{2.2.6}$$

The most used methods such as HAPMIX (Price et al., 2009), LAMP-LD (Churchhouse and Marchini, 2012) use the structure of the Hidden Markov Model (HMM) on ancestral haplotypes to model LD accuracy for genome wide data. Unfortunately, HAPMIX doesn't model admixture of more than two ancestral populations(Sundquist et al., 2008). The transition probability in equation 2.2.2 is also modelled differently depending coalescence parameters.

2.2.2 Hierarchies Hidden Markov Model. The Hierarchical Hidden Markov Model (HHMM) is used for a representation of possible emissions, and models a wide range of the correlation between haplotypes in order to capture the effects of LD at a large distance.

Suppose we have N populations $P = \{P_1, \cdots, P_N\}$ where each is represented by a set of n_p model individuals $P_p = \{a_{p1}, a_{p2}, \cdots, a_{pn_p}\}$. For each individual a_{pk}, we have SNP genotypes sampled at L loci spaced across the genome phased into two putative haplotypes $a_{pkh0} = \{a_{pk01}, a_{pk02}, \cdots, a_{pk0L}\}$ and $a_{pkh1} = \{a_{pk11}, a_{pk12}, \cdots, a_{pk1L}\}$ where, we can simplify by $a_{pkhi} = \{a_i\} \in \{A, C, G, T\}$ and $h \in \{0, 1\}$, where p, k, h and i represent populations, individuals, haplotypes and loci, respectively.

Let's consider S_{pkh} to be an emitting state for two haplotypes, there exist also the non-emitting states $\{I_p\}$ and $\{O_p\}$ for each population p, that serve as the primary means of considering the hidden variables are denoted by Y_i. The emission probability of the Hierarchical Hidden Markov Model (HHMM) is given by,

$$P(a_i = x|Y_i = S_{pkh}) = M(a_{pkhi}, x), \tag{2.2.7}$$

which can written as $M(x', x)$ where $x' = a_{pkhi}$ and $M$is the matrix has two entry x and x'.

For the initial emitting state, the model assumes each population has an equal probability given by

$$P(x_1 = S_{pkh}) = \frac{1}{2} N n_p. \tag{2.2.8}$$

$$S_{pkh} = \begin{cases} (1 - w_{pki})e^{-\tau_p R_i}, & \textit{or,} \\ w_{pki}e^{-\tau_p R_i} & \textit{for emitting state,} \\ 1 - e^{-\tau_p R_i} & \textit{non emitting state.} \end{cases} \tag{2.2.9}$$

Where τ_p is the reciprocal of the expected genetic length of a haploblocks inherited from population p and w_{pki} is the probability of a phasing switch error between loci i and $i+1$ for individual k in population p. R_i is the genetic distance between any two adjacent loci i and $i + 1$(Shai et al., 1998).

On the other hand, the transition can occur from S_{pkh} to non emitting state O_p and then to an I_p state with the probability of $N \times N$ matrix such that. $P(O_p \longrightarrow I_p) = A(p, p')$, and at the end back to an emitting haplotype state with uniform probability $\frac{1}{2} n_p$.

In addition,the most advantage of the HHMM is the flexibility in the level of analysis the model can handle the coalescence human model through hierarchical. The HHMM for pinpointing ancestry along the genome of admixed population is implemented in different programs including HAPAA(Sundquist et al., 2008) is high accuracy than SABER(Sundquist et al., 2008), when the number of generation of the admixture event increased (Shai et al., 1998). The HHMM estimates the local ancestry through the posteriors probabilities matrices using Forward-Backward or Expectation Maximum algorithm see in section 2.3.

2.2.3 Factorial Hidden Markov Model (FHMM). FHMM is a statistical model which can be used to solve the problem of local ancestry inference, along the genome of an admixed individual. In factorial hidden Markov model, the single chain of the hidden variable is replaced by a chain of hidden vector of independent factors. In this approach, the FHMM presentation allows to naturally decouple the state space into two parallel dynamic processes generating S^m and S^p where S^m and S^p present an unobserved state for maternal and paternal respectively (Bogdan et al., 2009; Jesse et al., 2013).

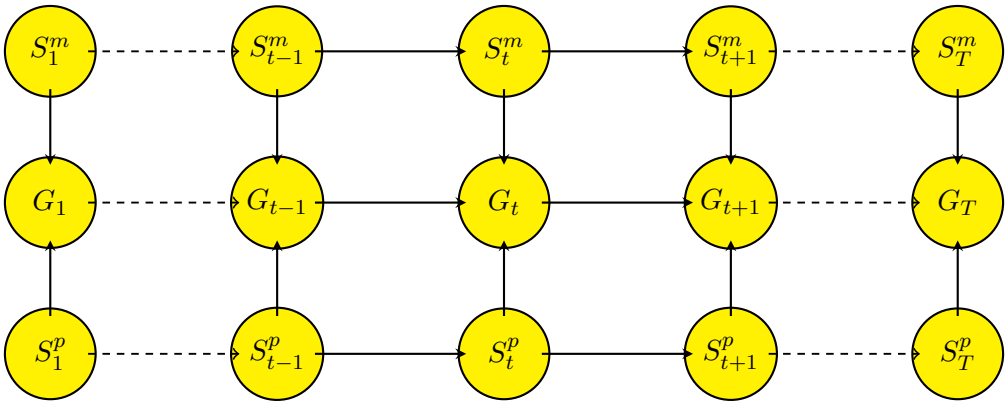


Figure 2.2: The graph shows the Factorial Hidden Markov model.

We can describe the factorial hidden Markov model in which a sequence of observation $\{G_t\}$ where $t = 1, \cdots \cdots, T$ across T genetic marker, is modelled by specifying a relation between observed and a sequence of hidden state $\{S_t\}$ and Markov transition structure linking the hidden states and observed state. The model assumes that two set of independence relation such that G_t is independent of all other observation and state given S_t and that S_t is independent of $S_1, \cdots, S_{t-2}$ given S_{t-1} with the first-order Markov property. By using the above independence we can found the joint probability for the sequence of the hidden state and the observed state and can be factorised as,

$$P(\{S_t, G_t\}) = P(S_1)P(G_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(G_t|S_t). \tag{2.2.10}$$

The transition probability can be governed through,

$$P(S_t|S_{t-1}) = P(q_t)P(S_t|S_{t-1}, q_t) + P(\overline{q}_t)P(S_t|S_{t-1}, \overline{q}_t). \tag{2.2.11}$$

Where q_t is defined as the event in which at least one past recombination occurred between position markers t and $t-1$, and $\overline{q}_t$ is defined as the complement event.

The recombination probabilities is given by,

$$P(S_t^k|S_{t-1}^k) = 1 - exp(-\phi(g.d_t)), k = 1, \cdots, K \quad , K \textit{ is the number of ancestries} \tag{2.2.12}$$

Where d_t is the genetic distance in Morgan between markers t and $t-1$, and $g+1$ the generations of the past recombination,$\phi(x)$ is a function of recombination rate.

We can deal with the local ancestry inference by calculating maximum posterior probabilities by referring on given genotypes by applying forward backward algorithm.

$$P(S_t^m = b_t, S_t^p = b_t'|G) \propto \alpha_t(b_t, b_t')\beta_t(b_t, b_t'). \tag{2.2.13}$$

We discussed the approximation of the posterior distribution in detail in Section 2.3.3. The main advantage of this approach in inferring the local ancestry is the improvement of the accuracy compared to other approach. In addition, the FHMM for pinpointing local ancestry along the adimxed population

is implemented different approach such as HAPAA, HAPMIX and ALLOY (Rodriguez et al., 2012) such that ALLOY shows highest accuracy to capture the parallel process producing the maternal and paternal admixed haplotype (Rodriguez et al., 2012). The FHMM estimates the local ancestry through the posteriors probabilities matrices using Forward-Backward or Expectation Maximum algorithm we will discussed in the detail in (Section 2.3).

2.3 Methods to Estimate the Posterior Distribution of Hidden Markov Models

2.3.1 Overview. In this part, we concentrate on the approximation of the maximum posterior distribution of the subset of the variables in the hidden Markov model as well as the parameter estimate through the Expectation Maximum EM algorithm. The parameter estimate can be obtained by the product of forward-backward parameter estimates divide with normalized.

2.3.2 Forward and Backward Algorithm for Standard Hidden Markov Model. Here we specify forward and backward recursion for computing the posterior probabilities of the hidden state FHMM. Using notation $\{Y_\tau\}_t^T$ to be the observation sequence defined as $Y_t, \cdots, Y_\tau$ and a given current parameter to estimate θ, $m \in \{1, \cdots, T\}$ Then,

$$\alpha_t = P((S_t^1, S_t^2, \cdots, S_t^m), \{Y_\tau\}_1^T|\theta). \tag{2.3.1}$$

Therefore,

$$\begin{aligned}
\alpha_t^0 &= P((S_t^1, S_t^2, \cdots, S_t^m), \{Y_\tau\}_1^{t-1}|\theta),\\
\alpha_t^1 &= P((S_{t-1}^1, S_t^2, \cdots, S_t^m), \{Y_\tau\}_1^{t-1}|\theta),\\
\alpha_t^1 &= P((S_{t-1}^1, S_{t-1}^2, \cdots, S_t^m), \{Y_\tau\}_1^{t-1}|\theta),\\
&\vdots\\
\alpha_{t-1}^m &= P((S_{t-1}^1, S_{t-1}^2, \cdots, S_{t-1}^m), \{Y_\tau\}_1^{t-1}|\theta) = \alpha_{t-1}.
\end{aligned} \tag{2.3.2}$$

Then we obtain the forward recursions

$$\alpha_t = P(Y_t|S_t^1, S_t^2, \cdots, S_t^m, \theta)\alpha_t^{(0)} \tag{2.3.3}$$

and

$$\alpha_t^{m-1} = \sum_{S_{t-1}^m} P(S_t^m|S_{t-1}^m)\alpha_t^{(m)}. \tag{2.3.4}$$

At the end of the forward recursions, the likelihood of the observation sequence is the sum of the K^m elements in α_T. Similarly, to obtain the backward recursion we define it as follows:

$$\beta_t = P(\{Y_\tau\}_{t-1}^T|S_t^1, S_t^2, \cdots, S_t^m, \theta), \tag{2.3.5}$$

$$\begin{aligned}
\beta_{t-1}^{m} &= P(\{Y_\tau\}_t^T | S_t^1, S_t^2, \cdots, S_t^m, \theta), \\
&\vdots \\
\beta_{t-1}^{1} &= P(\{Y_\tau\}_t^T | S_t^1, S_{t-1}^2, \cdots, S_t^m, \theta), \\
\beta_{t-1}^{0} &= P(\{Y_\tau\}_t^T | S_{t-1}^1, S_{t-1}^2, \cdots, S_t^m, \theta) = \beta_{t-1}.
\end{aligned} \tag{2.3.6}$$

Thus, we obtain,

$$\beta_{t-1}^{m} = P(\{Y_\tau\}_t^T | S_t^1, S_{t-1}^2, \cdots, S_t^m, \theta)\beta_t. \tag{2.3.7}$$

Hence,

$$\beta_{t-1}^{m-1} = \sum_{S_t^m} P(S_t^m | S_{t-1}^m)\beta_{t-1}^m. \tag{2.3.8}$$

Therefore, the posterior probability of the state at generation t is obtained by multiplying α_t and β_t and normalizing.

$$\gamma_t = P(S_t | \{Y_\tau\}_1^T, \theta) = \frac{\alpha_t.\beta_t}{\sum_{S_t} \alpha_t \beta_t}. \tag{2.3.9}$$

2.3.3 Forward-Backward Algorithm for FHMM. Let's consider FHMM S_t^m, S_t^p to be the hidden haplotype cluster membership drawn from a state space of the haplotype group A_t on maternal and paternal haplotype position t, respectively. Let $G_t \in \{0, 1, 2\}$, be the genotype observed state at the same markers position t. Furthermore, each haplotype group $b_t, b_t' \in A_t$ at location t is mapped to a single allele, denoted by $f(b_t) \in \{0, 1\}$. The vector of haplotype group memberships and genotype across all n markers position are denoted by $S^{m,p} = S_1^{(m,p)}, S_2^{(m,p)}, \cdots, S_n^{(m,p)}$, $G = G_1, G_2, \cdots\cdots, G_n$ respectively, where $t \in (1, 2, \cdots, n)$, m and p represent maternal and paternal, respectively. To infer local ancestry, we start by computing the posterior marginal given the sampled genotypes $P(S_t^m, S_t^p | G)$ by applying the forward and backward algorithm.

$$P(S_t^m = b_t, S_t^p = b_t' | G) \propto \alpha_t(b_t, b_t').\beta_t(b_t, b_t'), \tag{2.3.10}$$

where

$$\begin{aligned}
\alpha_t(b_t, b_t') &= P(G_1, \cdots, G_t | S_t^m = b_t, S_t^p = b_t'), \\
\beta_t(b_t, b_t') &= P(G_{t+1}, \cdots, G_n | S_t^m = b_t, S_t^p = b_t').
\end{aligned} \tag{2.3.11}$$

By using a naive recursive the complexity time for calculation of α, β $O(|A_1|^2 + \sum_{t=2}^{n} |A_{t-1}|^2 |A_t|^2)$. However, the dependency of FHMM allows for a more efficient recursive computation of α and β (Jesse

et al., 2013), where α is calculated in the forward direction in three steps as follows:

$$\begin{aligned}
\alpha^m_{t-1}(b_t, b'_{t-1}) &= \sum_{b_{t-1}\in A_{t-1}} \alpha_{t-1}(b_{t-1}, b'_{t-1}).P(S^m_t = b_t|S^m_{t-1} = b_{t-1}),\\
\alpha^p_{t-1}(b_t, b'_t) &= \sum_{b'_{t-1}\in A_{t-1}} \alpha_{t-1}(b_t, b'_{t-1}).P(S^p_t = b'_t|S^p_{t-1} = b'_{t-1}),\\
\alpha_t(b_t, b'_t) &= \alpha_{t-1}(b_t, b'_t).P(G_t|S^m_t = b_t, S^p_{t-1} = b'_{t-1}).
\end{aligned} \tag{2.3.12}$$

In the same manner β is calculated in the backward recursive in three steps as follows:

$$\begin{aligned}
\beta^p_t(b_t, b'_t) &= \beta_t(b_t, b'_t)P(G_t|S^m_t = a_t, S^p_t = b'_t),\\
\beta^m_t(b_{t-1}, b'_t) &= \sum_{b_t\in A_t} P(S^m_t = b_t|S^m_{t-1} = b'_{t-1})\beta^p(b_t, b'_t),\\
\beta_{t-1}(b_{t-1}, b'_{t-1}) &= \sum_{b_t\in A_t} P(S^p_t = b'_t|S^p_{t-1} = b'_{t-1})\beta^m_t(b_{t-1}, b'_t).
\end{aligned} \tag{2.3.13}$$

Thus, the computational cost of estimating β, on the maternal track is given $(|A_{t-1}|.|A_t|).|A_t|$ and the computation cost of estimate β paternal track is given by $(|A_{t-1}|.|A_{t-1}|).|A_t|$. It seems that a single forward step α_t is computed in complexity time equal to $|A_t|.|A_{t-1}|.(|A_t|+|A_{t-1}|)$. Therefore, the time has now become $O(|A_t|^2 + \sum_{t=2}^{n}|A_t|.|A_{t-1}|(|A_t| + |A_{t-1}|))$ (Jesse et al., 2013).

The emission probability $P(G_t|S^m_t, S^p_t)$ used in equations 2.3.12 and 2.3.13 is defined as

$$P(G_t|S^m_t = b_t, S^p_t = b'_t)\begin{cases} 1-2\epsilon, & f(b_t)+f(b'_t) = G_t,\\ \epsilon & \textit{otherwise}\end{cases} \tag{2.3.14}$$

where ϵ is the genetic error rate. Therefore, the maximal posterior is given by

$$\gamma_t(b_t, b'_t) = \frac{\alpha_t(b_t, b'_t)\beta(b_t, b'_t)}{\sum_{b_t, b'_t}\alpha_t(b_t, b'_t)}. \tag{2.3.15}$$

2.3.4 Expectation Maximization Algorithm or Classification Expectation Maximization algorithm. EM is also one of the approach used to approximate the maximum likelihood for the hidden variables. The EM is the method of iteration which follows two steps known as E-step and M- step up until its convergence(Sankararaman et al., 2008). The E-step refers to the calculation of the posterior probabilities of the parameter such $P(p_j, q_j|X_{i,j}, Z^t_{i,j})$ which can be computed using Baye's theorem. p_j, q_j are the probability of 1 occurring in the j^{th} SNPs in the first and second population, respectively, and $X_{i,j}, Z_{i,j}$ are the observed and unobserved variable, respectively. The M-step involves in the calculation of maximum likelihood which is given by,

$$\begin{aligned}
P(p_j, q_j|X_{i,j}, Z^t_{i,j}) &= log\,[P(Z_{i,1}|\alpha)] + I_{1,i}(Z_{i,1})\\
&\quad + \sum_{j=2}^{n}\{I_{j,i}(Z_{i,j}) + f_{i,j-1}(Z_{i,j-1}, Z_{i,j}, W_{i,j})\},
\end{aligned} \tag{2.3.16}$$

where,

$$f_{i,j-1}(Z_{i,j-1}, Z_{i,j}, W_{i,j}) = log\left[P(Z_{i,j}|Z_{i,j-1}, W_{i,j}, \alpha)\right] + log\left[P(W_{i,j}|\theta_j)\right] \tag{2.3.17}$$

and

$$I_{j,i}(Z_{i,j}) = \sum_{i=1}^{m}\sum_{j=1}^{m}\int\{log\left[P(X_{i,j}|Z_{i,j}, p_j, q_j)\right]P(p_j, q_j|X_{.,j}Z_{.,j}^{(t)})dp_j dq_j\}. \tag{2.3.18}$$

Then, the local ancestry is obtained by calculating the maximum posterior estimates of q and α which is given by

$$\textit{argmax}_{q,\alpha}\left[logPr(q, \alpha|X, p, \theta)\right] = \textit{argmax}_{q,\alpha}\left[logPr(X|q, p, \alpha, \theta)\right] \tag{2.3.19}$$

Therefore, the Expectation Maximum EM is the powerful method for local ancestry inference.

2.4 Discussion

We discussed different approaches of hidden Markov models to model the switch of the ancestry along the genome in admixed population. These different models of hidden Markov have some advantages and disadvantages depending on the parameter used to model biological coalescences. In spite all of these hidden Markov may handle infinite space the natural selection may happens during or after the admixture events; and that may masks the true ancestry in the present genotypes of admixed individual. In the next chapter, we will discuss the effect of natural selection of using the real data of multi-way admixed population for different implemented hidden Markov models.

3. Assessment of Local Ancestry Inference in Multi-way Admixed Populations

3.1 Introduction

From the previous studies, different methods have been suggested for inferring local ancestry in adxmed populations such as Africa-American, Latino and South Africa Coloured population. The results of these methods have been used in disease mapping and in understanding human history (Pasaniuc et al., 2013). However a major limitation is that inferring ancestry along the genome of a complex multi-way admixed populations such as the South African Coloured and Latino populations (Pasaniuc et al., 2013) is currently an unsolved problem (Pasaniuc et al., 2013; Chimusa et al., 2012). These current methods may reach high accuracy on average local ancestry using simulation, but in real data may at a particular regions obtain wrong deviations in the average local ancestry (e.g. regions in which the modelled ancestral population is unusually different from the true ancestral population due to the historical action of natural selection). In worse case, if these spurious deviations are present and differing in both affected and unaffected individuals from the same ethnic group, would lead to spurious mapping genes underlying ethnic difference in disease risk when applying disease scoring statistic.

The South African Coloured population used in this study, is known as complex admixed population from multiple source of populations (deWit et al., 2010; Chimusa et al., 2012). However, the second highest incidence of TB in the world is in the Western Cape in South Africa, particularly in this admixed South African Coloured (SAC) population (Moller et al., 2009; Moller and Hoal, 2010a,b). TB is a significant source of morbidity and mortality worldwide, particularly in developing countries.

In this chapter, we use the Tuberculosis (TB) data of the South Africa Coloured population to assess the effect of natural selection in both TB case and control individual of the admixed South African Coloured.

3.2 Materials and Methods

3.2.1 Data Description. The South African Coloured population used in this study, is known as complex admixed population with major genetic ancestral components, predominantly from Khoisan, Bantu-speaking Africans, European, Indian and east Asian contribution (deWit et al., 2010). This admixed population is located in the metropolitan area of Cape Town in the Western Cape Province in South Africa (deWit et al., 2010; Chimusa et al., 2012). The study samples were genotyped on the Affymetrix 500 K chip and SNP calling was done as described by De Wit et al. (deWit et al., 2010). The quality-control filters were initially applied to 797 cases and 91 controls in De Wit et al. (deWit et al., 2010), in this study we obtain $390,887$ SNPs for 733 individuals ($381,558$ autosomal SNPs, 642 cases and 91 controls; 406 males of which 361 are cases and 45 controls).

3.2.2 Methods. We used both LAMP-LD and WINPOP as they are known to currently be the most accurate local ancestry inference methods. We performed LAMP-LD and WINPOP in the TB case-control data of the South Africa Coloured population. Both LAMP-LD and WINPOP used the proxy ancestral haplotype data of the admixed South Africa Coloured. We selected the European (CEU), the

Yoruba (YRI), the Gujarati Indian (GHI), the Khoisan (KHS) and the Chinese (CHB) as proxy ancestral population of the South Africa Coloured (Chimusa et al., 2013).

In addition, both LAMP-LD and WINPOP estimate the ancestry along the genome of 733 individuals (642 affected TB and 91 unaffected individuals) in admixed South Africa Coloured. We applied a caster python script to computer the average local ancestry in the 733 admixed individuals, as well as in 642 TB case and 91 control. From the generated average local ancestry, we estimated region of unusual deviation in average local ancestry in all admixed individuals, as well as TB affected individuals and unaffected individuals. We compared these region of deviation in average local ancestry in case and control. We calculate the deviation of local ancestry of each ancestry as follows,

By considering a given the genome-wide ancestral proportion μ_k from ancestral populations $k \in \{1, \cdots, K\}$ in N samples of an admixed population, let X_k^{im} be the estimated local ancestry of individual i at genetic marker $m \in \{1, \cdots, M\}$, from the k^{th} ancestral population. We computed the deficiency or excess of ancestry, at each SNP using the estimated admixture proportion as a baseline. We thus normalized the deviations in the estimated local ancestry from ancestral population k at marker m in the admixed population by subtracting the mean and dividing by observed variance.

$$Z_k^m = \left(\frac{1}{N}\sum X_k^{im}\right) - \mu_k = \overline{X}_k^m - \mu_k, \tag{3.2.1}$$

where $\overline{X}_k^m$ is the average local ancestry at SNP m. Z_k^m can be approximated as a normal distribution under mean 0 and variance 1, derived from the distribution of $\sum X_k^{im}$ values among the N individuals. Thus, we fit a chi-square as follows,

$$\chi_k^m = \frac{(X_k^m)^2}{Var(X_k^{im})} \tag{3.2.2}$$

is χ^2 with 1 degree of freedom.

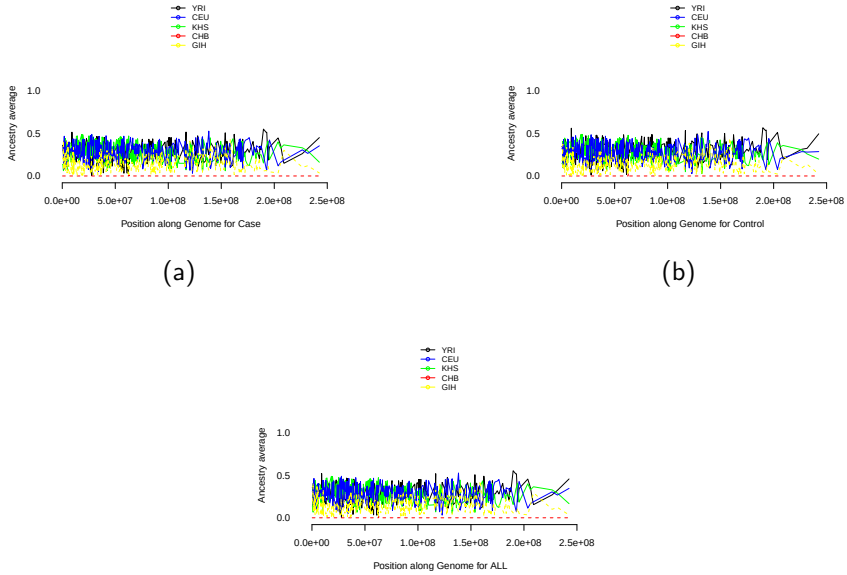
3.3 Result and Discussion

To assess the effect of natural selection in pinpointing local ancestry along the genome of the admixed population for both LAMP-LD and WINPOP using the proxy ancestral haplotype data within the genotype data of the admixed South Africa Coloured population.

We estimated for 733 individuals the mixture of European (CEU), Bantu (YRI), Gujarati (GHI), Khohisan (KHS) and Chinese (CHB). In order to assess the accuracy of both LAMP-LD and WINPOP, we plotted estimated average local ancestry versus the physical position along the genome for case which is illustrated in (Figure 3.1a) for LAMP-LD and in (Figure 3.2a) for WINPOP. The plot of the estimated average local ancestry for control for both LAMP-LD and WINPOP, it is illustrated in (Figure 3.1b) and (Figure 3.2b) as well as for all populations for both two approaches are shown in (Figure 3.1c) and (Figure 3.2c).

In addition, by comparing the proportion of ancestral population in the South Africa Coloured population. We observed that the Chinese (CHB) has lower proportion of 1% for all population as well as for case and control using LAMP-LD while Yoruba has high proportion of 29%, 29% and 30% for all populations, case and control, respectively (see Table 3.1). For WINPOP, there is a little bit change of proportion of ancestral population contribution in the South Africa Coloured population, for example the Chinese (CHB) who has a lower proportion of 1.4%, 15% and 19% for all South African Coloured individuals, case and control individuals, respectively. Similarly, the ancestral population has higher proportion in the South Africa Coloured population is the Yoruba (YRI) who has 49% and 24% for all population and case, respectively and the European (CEU) has higher proportion for control of 26%. Also the proportions of other ancestral population are illustrated in (Table 3.1) in details.

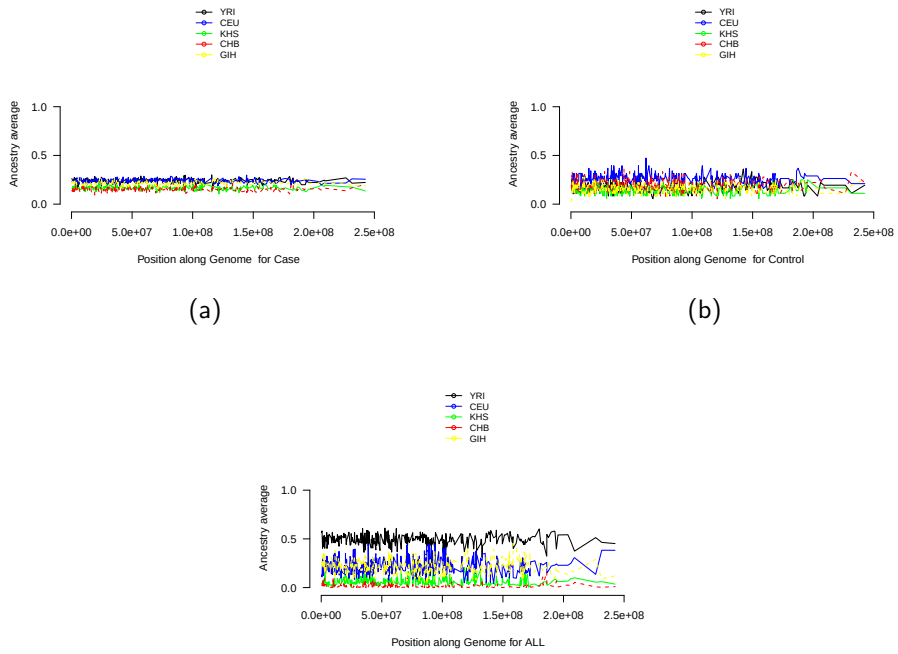
By observing the variation of the proportion of the ancestral population in the South Africa Coloured population, we can see that, LAMP-LD are similar magnitude of proportion from all South African Coloured individuals as well as for case and control individuals. But for WINPOP, we have observed different in the ancestral proportion between case and control individuals. From previously published ancestral proportion of this admixed population (Chimusa et al., 2012; deWit et al., 2010), we note that LAMP-LD produced better results than WINPOP. In addition, LAMP-LD is more stable in the variation of the proportion of the average local ancestry for case-control as well as in the case of the entire sample of the South African Coloured population.



(a) (b)

(c)

Figure 3.1: Plot of average local ancestry using LAMP-LD in figure 3.1a for case, 3.1b for control and 3.1c for ALL



(a) (b)

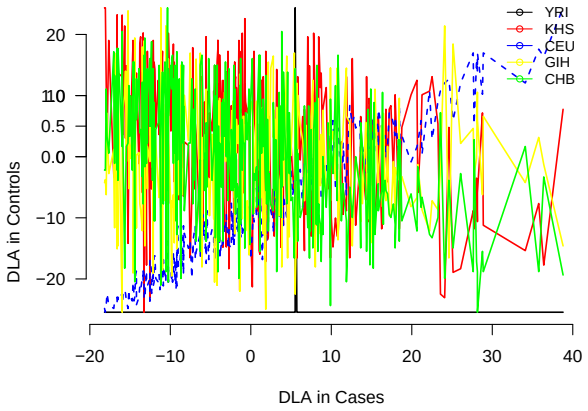
(c)

Figure 3.2: Plot of average local ancestry using WINPOP in figure 3.2a for case, 3.2b for control and 3.2c for all population

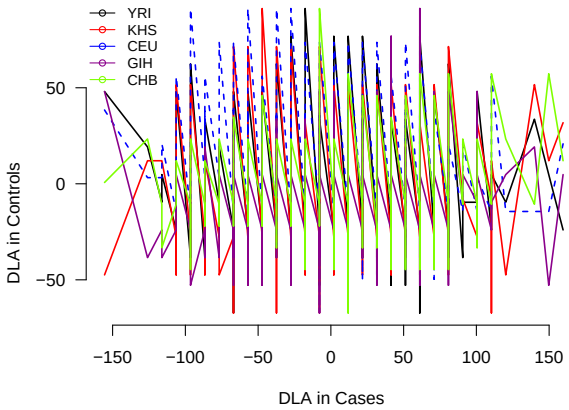
Figure 3.3 illustrates the deviation in the average local ancestry for both LAMP-LD and WINPOP. The plot of cases versus controls shows spurious deviation in average local ancestry even if they come from the same populations. The case and control must have the same magnitude of deviation in average local ancestry. Since all population come from the same populations. Thus these approaches are not able to fit the effect of natural selection that may occurs during or after admixture event.

At least, from the result in Figure 3.3a obtained from LAMP-LD, only the European CEU ancestry was well modelled as there is similar magnitude of the deviation in average local ancestry between cases and controls.

This implies that European CEU has recently been able to adapt in the new environment caused by natural selection effect. Since the model LAMP-LD has applied the structure of linkage disequilibrium on the represented ancestral populations.



(a) Deviation in local ancestry using LAMP LD method



(b) Deviation in local ancestry using WINPOP method

Figure 3.3: The plot show the deviations in local ancestry

Methods	LAMP-LD					WINPOP				
Population	KHS	YRI	CEU	CHB	GIH	KHS	YRI	CEU	CHB	GIH
ALL	28%($\pm$0.1)	29%($\pm$0.095)	28%($\pm$0.0974)	1%($\pm$0.00047)	14%($\pm$0.857)	6%($\pm$0.385)	49%($\pm$0.049)	21%($\pm$0.079)	1.4%($\pm$0.018)	21.9%($\pm$0.072)
Case	28%($\pm$0.1)	29%($\pm$0.095)	28%($\pm$0.097)	1%($\pm$0.00044)	14%($\pm$0.86)	16%($\pm$0.096)	24%($\pm$0.02)	24%($\pm$0.020)	15%($\pm$0.02)	21%($\pm$0.019)
Control	28%($\pm$0.1)	30%($\pm$0.098)	27%($\pm$0.098)	1%($\pm$0.00077)	14%($\pm$0.868)	15%($\pm$0.041)	18%($\pm$0.049)	26%($\pm$0.05)	19%($\pm$0.054)	16%($\pm$0.d046)

Table 3.1: Table shows the proportion of each ancestral population of 733 admixed South Africa Coloured individuals. In addition, the table displays also the Genome-wide ancestral proportion of 642 Tb case and 91 control individuals of the South Africa Coloured population. The value in parentheses is the standard deviation for each ancestral proportion.

4. Conclusion and Future Work

In this project, we discussed in details various hidden Markov models of switches in ancestry between consecutive windows of loci and the fit model parameters using the model-Markov chain Monte Carlo sampling; forward-backward algorithm and the Expectation Maximization algorithm. We have assessed the effect of natural selection in different implemented hidden Markov Models using the Tuberculosis case-control data of the South African Coloured population. We have examined spurious deviations in the average local of TB case and control of the admixed South African Coloured population using most current known accurate local ancestry inference methods include WINPOP and LAMPLD.

We have observed that both local ancestry inference methods include WINPOP and LAMPLD have yielded several regions of spurious deviation in the average local ancestry in both TB case and control individuals of the South African Coloured population. Our results are consistent with previous finding in Chimusa et al. 2013, suggesting that pinpointing ancestry along the genome of a multi-way admixed population such as the Latino and the South African Coloured population is currently a major challenge and unsolved problem. A robust probabilistic approach that can model the natural selection in pinpointing ancestry along the genome of an admixed individual is needed, and that will be a breakthrough for conducting a proper admixture mapping, admixture association or applying in disease scoring statistic using admixed populations. In particular, the application of local ancestry will be crucial for complex diseases that show difference in prevalence among different ethnic groups.

Acknowledgements

I would like to thank almighty God, Source of my life for giving me power and strength of thinking and writing this project.

It is a great pleasure to thank my supervisor Dr Emile Chimusa Rugamika for his assistance and supervision which made this project possible.

To All AIMS staff, for their technique support and advice

To my family, friends for their support and encouragement.

References

L. Alkes. Chromosomal segments in admixed individualsand inference of local ancestry. Available from http://www.hsph.harvard.edu/alkes-price/files/2012/10/ashg_price_110310_localanc.pdf, 2010.

L. P. Alkes, T. Arti, P. Nick, C. B. Kathleen, R. Nicholas, R. Ingo, H. B. Terri, M. Rasika, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PloS Genetics*, 5:6, 2009.

Y. Baran, P. Bogdan, S. Sankararaman, G. Dara, C. Gignoux, C. Celeste, W. Torgerson, R. Chapela, J. G, C. Pedro, J. Rodriguez-Santana, E. Gonzalez, and E. Eran. Fast and accurate inference of local ancestry in latino populations. *Bioinformatics. 28, 1359-1367*, 2012.

P. Bogdan, K. Justin, and M. Ion. Imputation-based local ancestry inference in admixed populations. *Springer-verlay Berlin*, 5542:221– 233, 2009.

M. Brenna, R. Laura, G. Simon, W. Wei, B. Abra, K. Jake, F.-Z. Karima, A. Pierre, M.-E. Andres, B. Jaume, and D. Carlos. Genomic ancestry of north africans supports back-to-africa migrations. *PloS Genetics*, 5:6, 2012.

E. Chimusa, Z. Noah, D. Michelle, M. Marlo, N. Mulder, A. Price, and E. Hoal. Genome-wide association study of ancestry-specific tb risk in the south african coloured population. *Hum.Mol. Genet. doi:10.1093/hmg/ddt462*, 2012.

E. Chimusa, M. Daya, M. Möller, R. Ramesar, B. Henn, P. Helden, N. Mulder, and E. Hoal. Determining ancestry proportions in complex admixture scenarios in south africa using a novel proxy ancestry selection method. *PLoS ONE 8(9): e73971.doi:10.1371/journal.pone.0073971*, 2013.

E. Christopher, A. Stephen, E. Peter, and J. Robert. Detecting and measuring genetic differentiation. *Crust Issues*, 19:31–39, 2011.

C. Churchhouse and J. Marchini. Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet. Epidemiology. 37, 1-12*, 2012.

M. E. David and R. C. Lon. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *American Journal of Human Genetics*, 76:681–687, 2005.

E. deWit, W. Delport, R. Chimusa, A. Meintjes, M. Moller, P. Helden, C. Seoighe, and V. Hoal. Genome-wide analysis of the structure of the south African Coloured population in the western Cape. *Hum. Genet. 128, 15-53*, 2010.

DF.Conrad, M.Jakobsson, G.Coop, X.Wen, JD.Wall, N. Rosenberg, and J. Pritchard. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet. 38, 1251-1260*, 2010.

J. W. Dower. Readings compiled for history. 21:1–4, 1991.

J. P. Esteban, M. Amy, A. Joshua, M. Jeremy, A. B. Mark, C. Richard, F. Terrence, B. A. David, R. Deka, E. F. Robert, and D. S. Mark. Estimating african american admixture proportions by use of population specific alleles. *The American Society of Human Genetics*, 63:1839– 1851, 1998.

D. Evans and L. Cardon. A comparison of disequilibrium patterns and estimated population recombination rates across multiple populations. *Am. J. Hum. Genet. 76, 681-687*, 2005.

J. F.J and C.H.kuo. bottlesim: a bottleneck simulation program for long-livedspecies with overlapping generations. *C214 Life Sciences*, 2003.

K. E. H. Gad, S. Sriram, and P. Bogdan. Inference of locus-specific ancestry in closedly related populations. *ISMB*, 20:i213–i221, 2009.

D. Goldstein and M. Weale. Population genomics: linkage disequilibrium holds the key. *Curr. Biol. 11(14), R576-9*, 2001.

S. Gravel. Population genetics models of local ancestry. *The Genetics Society of America*, 2012.

H. Halder and S. Shriver. Measuring and using admixture to study the genetics of complex diseases. *Hum. Genomics. 1, 52-62*, 2003.

B. Henn, L. Botigue, S. Gravel, W. Wang, A. Brisbin, J. Byrnes, K. Fadhlaoui-Zid, P. Zalloua, A. Moreno, J. Bertranpetit, C. Bustamante, and D. Comas. Genomic ancestry of north africans supports back to africa migrations. *Nat. Comm. 3 (1143) 2140*, 2012.

M. Jesse, B. Sivan, E. Megan, and S. Batzoglou. Ancestry inference in complex admixtures via variable-length markov chain linkage models. *Journal of Computation Biology*, 25:199–211, 2013.

H. r. John and M. H. Rosalind. Population genetics of modern human evolution. *Macmillan publishind Ltd*, 2001.

C. Kristin, L. Kruglyak, and M. Seielstad. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genet. 3, 299-309*, 2002.

S. Kyung-Ah and P. Eric. Spectrum: joint bayesian inference of population structure and recombination events. *ISMB/ECCB*, 23:i479–i489, 2007.

D. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genet. 8(1),e1002453*, 2010.

Y. Liu, T. Nyunoya, S. Leng, S. Belinsky, Y. Tesfaigzi, and S. Bruse. Software and methods for estimating genetic ancestry human population. *Hum Genomics. 7(1), 1*, 2013.

C.-S. L.Luca and W. F. marcus. The application of molecular genetic approaches to the study of human evolution. *natural genetic supplemental*, 33:266–275, 2003.

K. Lohmueller, A. Albrechtsen, Y. Li, S. Y. Kim, T. Korneliussen, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, A. F. Feder, N. Grarup, T. Jørgensen, T. Jiang, D. R. Witte, A. Sandbæk, I. Hellmann, T. Lauritzen, T. Hansen, O. Pedersen, J. Wang, and R. Nielsen. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet 7(10), e1002326*, 2011.

B. Mary, McEachern, H. V. Drik, H. chris, m. Bernie, and M. John. Bottlenecks and rescue effects in a fluctuating population of golden mantled ground squirrls (spermorphilus lateralis). *Conserv genet*, 12:285–296, 2010.

P. McKeigue. Prospects for admixture mapping of complex traits. *Am. J. Hum. Genet. 76(1), 1-7*, 2005.

M. Moller and E. Hoal. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis. 90, 71-83*, 2010a.

M. Moller and E. Hoal. Past, present and future directions in human genetic susceptibility to tuberculosis. *FEMS Immunol Med Microbiol. 58, 3-26*, 2010b.

M. Moller, A. Nebel, R. Valentonyte, S. S. Helden van, and E. Hoal. Investigation of chromosome 17 candidate genes in susceptibility to tb in a south African population. *Tuberculosis. 89, 189-194*, 2009.

G. Montana and J. Pritchard. Statistical tests for admixture mapping with case-control and case-only data. *Am. J. Hum. Genet. 75(5), 771-789*, 2004.

P. Nick, M. Priya, L. Yontao, M. Swapan, R. Nadin, T. G. Yiping, Zhan, and W. D. R. Teresa. Ancient admixture in human history. *genetics*, 192:1065–1093, 2012.

B. Pasaniuc, S. Sankararaman, G. Kimmel, and E. Halperin. Inference of locus-specific ancestry in closely related population. *Bioinformatics. 25, i213-i221*, 2009.

B. Pasaniuc, N. Zaitlen, G. Lettre, G. Chen, A. Tandon, W. L. Kao, I. Ruczinski, M. Fornage, D. Siscovick, X. Zhu, E. Larkin, L. Lange, A. Cupples, Q. Yang, E. Akylbekova, S. Musani, J. Divers, J. Mychaleckyj, M. Li, G. Papanicolaou, R. Millikan, C. Ambrosone, E. John, L. Bernstein, W. Zheng, J. Hu, R. Ziegler, S. Nyante, E. Bandera, S. Ingles, M. Press, S. Chanock, S. Deming, J. Rodriguez-Gil, C. Palmer, S. Buxbaum, L. Ekunwe, J. Hirschhorn, B. Henderson, S. Myers, C. Haiman, D. Reich, N. Patterson, J. Wilson, and A. Price. Enhanced statistical tests for gwas in admixed populations: Assessment using African Americans from CARe and a breast cancer consortium. *PLoS Genet.7(4), e1001371*, 2011.

B. Pasaniuc, S. Sankararaman, D. Torgerson, C. Gignoux, N. Zaitlen, C. Eng, W. Rodriguez-Cintron, R. Chapela, J. Ford, P. Avila, J. Rodriguez-Santana, G. Chen, L. Le, B. Henderson, D. Reich, C. Haiman, B. Gonzàlez, and E. Halperin. Analysis of latino populations from gala and mec studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics. 29(11):1407-15. doi: 10.1093/bioinformatics/btt166*, 2013.

A. Price, A. Tandon, N. Patterson, K. Barnes, N. Rafaels, I. Ruczinski, T. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *Plos Genet. 5, e1000519*, 2009.

S. Pritchard, M. Stephens, and M. Donnelly. Inference of population structure using multi-locus genotype data. *Am. J. Hum. Genet. 155, 945-959*, 2002.

H. Qin, N. Morris, S. Kang, M. Li, B. Tayo, H. Lyon, J. Hirschhorn, R. Cooper, and X. Zhu. Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics. 26, 2961-2968*, 2010.

C. Ranajit and M. W. Kennethe. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Crust Issues*, 19:31–39, 1988.

D. Redden, J. Divers, L. Vaughan, H. Tiwari, T. Beasley, J. Fernández, R. Kimberly, Feng, M. Padilla, N. Liu, M. Miller, and D. Allison. Regional admixture mapping and structured association testing: conceptual unification and an extensible general linear model. *PLoS Genet. 2(8), e137*, 2006.

D. Reich, N. Patterson, P. Jager, and G. McDonald. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat. Genet. 37, 1113-1118*, 2005.

J. Relethford and R. Harding. Population genetics of modern human evolution. *Encyclopedia of Life Sciences*, 2001.

J. Rodriguez, S. Bercovici, M. Elmore, and S. Batzoglou. Ancestry inference in complex admixtures via variable-length markov chain linkage models. *J. Comput. Biol. 20(3)*, 2012.

N. Rosenberg and J. Pritchard. Genetics structure of human populations. *Science. 298, 2381-2385*, 2008.

N. Rosenberg, J. Pritchard, J. Weber, H. Cann, K. Kidd, L. Zhivotovsky, and M. Feldman. Genetic structure of human population. *Science journal*, 289:298(5602):2381–5, 2002.

N. Rosenberg, L. Li, R. Ward, and J. Pritchard. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet. 73, 1402-1422*, 2003.

S. Sagiv, K. Jane, K. Mark, Y. Benjamin, and D. Ariel. Linkage disequilibrium patterns of the human genome across populations. *Human Molecul*, 12:771–776, 2003.

S. Sankararaman, G. Kimmel, E. Halperin, and M. Jordan. On the inference of ancestries in admixed populations. *Genome Res. 18(4), 668-675*, 2008.

M. Seldin, B. Pasaniuc, and A. Price. New approaches to disease mapping in admixed populations. *Nat Rev Genet. 36, S21-S27*, 2011.

F. Shai, S. Yoram, and T. Naftali. The hierarchiacl hidden markov model:analysis and application. *machine learning. 32, 41-62*, 1998.

S. S. S. Sriram, K. Gad, and H. Eran. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82:290–303, 2008.

J. Stephens, D. Briscoe, and S. O. SJ. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nature Rev. Genet. 6, 623-632*, 2005.

A. Sundquist, E. Fratkin, B. C. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using hapaa. *Genome Res. 18, 676-682*, 2008.

H. Tang, M. Coram, P. Wang, X. Zhu, and N. Risch. Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet. 79, 1-12*, 2006.

S. Tishkoff and K. Kidd. Implications of biogeography of human populations for 'race' and medicine. *Nat Rev Genet. 36, S21-S27*, 2004.

K. Verhoeven, M. Macel, M. L. Wolfe, and A. Biere. Population admixture, biological invasions and the balance between local adaptation and inbreeding depression. *Proc. R. Soc. B. 278, 2-8*, 2010.

B. Weir. linkage disequilibrium and association mapping. *Ann. Rev. of Genomics and Hum. Genet. 9, 129-142*, 2008.

B. Weir and C. Cockerham. Estimating f-statistics for the analysis of population structure. *Evolution. 38, 1358-1370*, 1984.

Z. Xiaofeng, T. Hua, , and R. Neil. Admixture mapping and the role of population structure for localizing disease genes. *Advances in Genetics*, 60:DOI: 10.1016/S0065–2660(07)00419–1, 2008.